Course: CMSC 424 – Database design
Instructor: Mihai Pop
Times: TuTh 11:00-12:15
Location: CSIC 1121

Office hours:
  Wed, 11-12, AVW 3223
  and by appointment

TA: Sharath Srinivas
TA office hours: TBA

Class website:
http://www.cbcb.umd.edu/confcour/CMSC424.shtml
Textbook: Database systems concepts.
Silberschatz, Korth, Sudarshan

Note: Lectures trump book
Both owned by Larry Ellison, CEO of Oracle

It pays to know databases!
Workload

• Exams: 2 midterms, 1 final
• Projects: 1 group programming project - build a database that does something cool (TBA)
• Homeworks: ~4 homeworks throughout the semester (some include SQL programming)

• Grading:
  – homeworks 10%
  – midterms 25%
  – final 25%
  – project 40%
Policies

• Attendance - follow University policy
  – you must claim excused absences in writing
  – written documentation of illness is required (from Dr. not yourselves)
  – if possible inform me prior to the class you will skip

• Disabilities
  – must inform me during the first 2 weeks of the semester if special accommodations necessary
  – request letter from Office of Disability Support Services

• General – communication is key
  – talk to me about any issues whether covered or not by University policies
Academic Honesty

http://www.studenthonorcouncil.umd.edu/code.html

• No cheating on homeworks/projects/exams
• No making up data/results
• No copying of other people’s code
• You can work together on homeworks/projects but WRITE THE ANSWER BY YOURSELF

I pledge on my honor that I have not given or received any unauthorized assistance on this examination.
Addl. Rules

• NO EXCUSE FOR CHEATING!

• NO LAPTOPS IN CLASS!
Why go through all this?

• Database administrators are paid well
• Databases are everywhere (i.e. lots of job opportunities)
  – E.g. Google
  – at the doctor's office
  – payroll systems
  – on Wall Street
  – government (e.g. CIA)
  – scientific data
• Database research offers many exciting opportunities
  – Internet technologies
  – handling huge amounts of data
  – etc.
Databases in the wild

• Database assembles US warnings of Saddam threat – Reuters (1/23/2008)
  – can search by keywords
  – summarizes statistics
  – assembled from a number of sources
  – manual curation/entry

• Google
  – database of searches (google trends)
  – database of emails (gmail)
  – database of publications (google scholar)
  – ...
  – privacy issues

• Bio-medical databases
  – doctor's office, lab providers, hospitals, research institutes
  – insurance companies
  – who/how/when/how much information shared?
Motivation: Data Overload

• Much more is produced every day

Wal-mart: 583 terabytes of sales and inventory data
Adds a billion rows every day
“we know how many 2.4 ounces of tubes of toothpastes sold yesterday and what was sold with them”

Yes we can do it; is there any point to it?

[[“library of congress --> 20 TBs”]]
Motivation: Data Overload

• Much more is produced every day

Neilsen Media Research: 20 GB a day; total 80-100 TB
From where ???
  12000 households or personal meters
  Extending to iPods and TiVos in recent years

Is there a point beyond telling you what great TV shows you are missing ?
Motivation: Data Overload

- Scientific data is literally astronomical on scale

  Sanger Center – 22 TB doubling every 10 months
  GenBank – 252 GB
  Trace Archive – 1.8 billion records (> 2 TB)

  New technologies – bwtn. 1TB and 100TB / day

Shameless plug: CMSC 423: bioinformatic algorithms, databases and tools. Fall 2008

Sloan Digital Sky Survey – 15 TB
Motivation: Data Overload

• Automatically generated data through instrumentation

“Britain to log vehicle movements through cameras. 35 million reads per day.”

Wireless sensor networks are becoming ubiquitous. RFID: Possible to track every single piece of product throughout its life (Gillette boycott)
Motivation: Data Overload

- How do we do *anything* with this data?
- Where and how do we store it?
  - Disks are doubling every 18 months or so -- not enough
- How do we search through it?
  - Text search?
  - “how much time from here to pittsburgh if I start at 2pm?”
  - Data is there; more will be soon (live traffic data)
Motivation: Data Overload

• What if the disks crash?
  ● Very common, especially if we are talking about 1000’s of disks storing a single system

• Speed!!
  – Imagine a bank and millions of ATMs
    ● How much time does it take you to do a withdrawal?
    ● The data is not local
  – How do we ensure “correctness”?
    ● Can’t have money disappearing
    ● Harder than you might think
DBMS to the Rescue

• Provide a systematic way to answer most of these questions…
• Aim is to allow easy management of data
  – Store it
  – Update it
  – Query it
• Massively successful for *structured* data
  – What do I mean by that?
Structured vs Unstructured

A lot of the data we encounter is *structured*
- Some have very simple structures
- E.g. Data that can be represented in tabular forms
- Significantly easier to deal with
- We will actually focus on such data for much of the class

<table>
<thead>
<tr>
<th>Account</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bname</strong></td>
<td><strong>cname</strong></td>
</tr>
<tr>
<td>A-101</td>
<td>Jones</td>
</tr>
<tr>
<td>A-215</td>
<td>Smith</td>
</tr>
<tr>
<td>Downtown</td>
<td>Hayes</td>
</tr>
<tr>
<td>A-102</td>
<td>Curry</td>
</tr>
<tr>
<td>Mianus</td>
<td>Lindsay</td>
</tr>
<tr>
<td>A-305</td>
<td></td>
</tr>
<tr>
<td>Perry</td>
<td></td>
</tr>
<tr>
<td>R.H</td>
<td></td>
</tr>
</tbody>
</table>
Structured vs Unstructured

• Some data has a little more complicated structure
  – E.g graph structures
    • Map data, social networks data, the web link structure etc
  – In many cases, can convert to tabular forms (for storing)
  – Slightly harder to deal with
    • Queries require dealing with the graph structure
Collaborations Graph

Query: Find my Erdos Number.
Structured vs Unstructured

- Increasing amount of data in a *semi-structured* format
  - XML – Self-describing tags
  - Complicates a lot of things
  - We will discuss this toward the end
Structured vs Unstructured

• A huge amount of data is unfortunately *unstructured*
  – Books, WWW
  – Amenable to pretty much only *text search*
    • Information Retrieval deals with this topic
  – What about Google?
    • Google is actually successful because it uses the structure
DBMS to the Rescue

• Provide a systematic way to answer most of these questions…
  – … for structured data
  – … increasing for semi-structured data
    • XML database systems have been coming up

• Solving the same problems for truly unstructured data remains an open problem
  – Much research in Information Retrieval community
  – think YouTube (what does a query for “train” retrieve)
DBMS to the Rescue

- They are everywhere!!
- Enterprises
  - Banks, airlines, universities
- Internet
  - Searchsystems.net lists 35568 public records DBs
  - Amazon, Ebay, IMDB
- Blogs, social networks…
- Your computer (emails especially)
- …
Out of scope…

• How do we guarantee the data will be there 10 years from now?
  – Much harder than you might think

• Privacy and security !!!
  – Every other day we see some database leaked on the web

• New kinds of data
  – Scientific/biological, Image, Audio/Video, Sensor data etc

• Interesting research challenges!
What we will cover...

• representing information
  – data modeling

• languages and systems for querying data
  – complex queries & query semantics
  – over massive data sets

• concurrency control for data manipulation
  – controlling concurrent access
  – ensuring transactional semantics

• reliable data storage
  – maintain data semantics even if you pull the plug
What we will cover…

• We will see…
  – Algorithms and cost analyses
  – System architecture and implementation
  – Resource management and scheduling
  – Computer language design, semantics and optimization
  – Applications of AI topics including logic and planning
  – Statistical modeling of data
What we will cover…

• We will mainly discuss structured data
  – That can be represented in tabular forms (*called Relational data*)
  – We will spend some time on XML

• Still the biggest and most important business
  – Well defined problem with really good solutions that work
    • Contrast XQuery for XML vs SQL for relational
  – Solid technological foundations

• Many of the basic techniques however are directly applicable
  – E.g. reliable data storage etc

• Many other data management problems you will encounter can be solved by extending these techniques